**Topological Bias and Inconsistency of Maximum Likelihood Using Wrong Models**

*William J. Bruno** *and Aaron L. Halpern*†
*Theoretical Biology and Biophysics, Los Alamos National Laboratory; †Health Sciences Center, Department of Molecular Genetics and Microbiology, University of New Mexico.
Corresponding Author:
William J. Bruno, Mail Stop K-710, Los Alamos, NM 87545; (505) 665-3802; fax: (505) 665-3493; email: billb@lanl.gov
**keywords:** long branches attract, most likely unresolved tree, positively misleading.
**running head:** Letter to the Editor
**abbreviations:** ML: maximum likelihood, LBA long branches attract, MLUT most likely unresolved tree.

Ziheng Yang (1997) presented interesting and surprising results suggesting that maximum likelihood (ML) using a simple model could do a better job of phylogenetic reconstruction than ML using the correct model. We believe Yang's results and others that we present here can best be understood in terms of bias in the choice of topology resulting from using ML with an incorrect model.

Yang simulated sequences related by various 4-taxon trees, using the Jukes-Cantor (JC) model with gamma-distributed site-to-site rate variation (JC+G), and a shape parameter $\alpha = 0.2$. He then used ML to reconstruct the phylogeny, assuming either the true, JC+G model with $\alpha = 0.2$, or the simpler but false JC model. In repeated simulations, the false model reconstructed the correct topology more often than the true model did for three of the five trees Yang tested (namely, trees B, D, and E). With a fourth tree (Yang's tree A), the two models performed equally well. On the fifth tree (C), the true model performed dramatically better. Trees generated randomly by a coalescent process were also reconstructed correctly more often using the false model. Yang's results held over a wide range of sequence lengths, suggesting ML with the true model to be less efficient than ML with the false model. This brought into question the theorem that ML, using the correct model, is asymptotically most efficient. Yang suggested one possible explanation for his results: that the proof of asymptotic efficiency of ML might not apply to the discrimination of nonnested models (e.g., different topologies).

We put forward here an alternative explanation: that Yang's results do not reflect an overall increase in efficiency when using the simpler model of sequence evolution, but rather reflect comparison of a biased, sometimes inconsistent method (ML with false model) to a nearly unbiased, consistent method (ML with true model). As we will show, it appears that most of the trees Yang simulated happen to have topologies favored by the bias of the false model.

Others have previously pointed out that false, oversimplified models cause the number of substitutions to be underestimated. The undercounting of substitution events occurs at all distances, but becomes more severe for more divergent sequences. This will lead to underestimation of longer pairwise distances relative to shorter distances (Gojobori, Ishii and Nei 1982), and also biased or inconsistent estimation of topology (Felsenstein 1978; Huelsenbeck

and Hillis 1993; Kuhner and Felsenstein 1994; Gaut and Lewis 1995; Swofford, Olsen, Waddell and Hillis 1996, page 442).

Perhaps the best known instance of this involves parsimony. Parsimony, by neglecting back-substitutions, also underestimates substitutions more severely with increasing divergence. Felsenstein noted that maximum parsimony has a topological bias (and inconsistency) in which long branches attract (LBA). In the so-called "Felsenstein zone"—characterized by trees with long branches on opposite ends of the internal branch—LBA favors the wrong topology and parsimony is inconsistent. However, one can also observe an "anti-Felsenstein zone"—characterized by trees with long branches on the same end of the internal branch—where LBA favors the correct tree and parsimony generally returns the correct topology more often than ML (Russo, Takezaki and Nei 1996). Thus, biases can lead to either greater or reduced efficiency, depending on the particular tree under consideration.

ML can show similar biases if the assumed model of evolution is overly simple (e.g., overly homogeneous) for a given data set (Kuhner and Felsenstein 1994; Gaut and Lewis 1995). Like parsimony, ML using an overly simple model, such as Yang's false model which ignores rate heterogeneity, will underestimate homoplasy (sharing of character states due to back replacements or convergence). As one consequence, ML using such a model is biased and inconsistent for the simple problem of finding the distance between two taxa. This effect can also lead to an LBA bias, for essentially the same reasons as in parsimony: when the overall number of substitutions is underestimated, a better fit (higher likelihood for ML, fewer changes for parsimony) results if shared characters between the taxa on long branches are interpreted as the result of shared evolution.

Such an effect seems to readily explain the results on Yang's trees B and C. On tree B, in which the long branches are joined, the simple model performs better than the true model, whereas on tree C, in which the long branches do not join, the simple model performs much worse than the true model. We present in Table 1 additional results that support our interpretation: tree B1, which is similar to tree B except that the long branches do not join, caused the simple model to perform worse than the true model. Conversely, tree C1, which resembles tree C except that the long branches join, caused the simple model to perform better.

Moreover, for the other trees on which the simple model performed better (trees D and E), we present similar trees (D1 and E1) on which it performs worse. Because trees D and E do not have a unique pair of longest external branches, one cannot directly invoke the LBA bias to explain Yang's results on these trees. However, we believe that essentially a single bias phenomenon is at work in all of the four-taxon trees we have studied. We describe below a possible connection between LBA and the biases in trees D and E.

Quantitative measures of tree shape bias have been described elsewhere (Kuhner and Felsenstein 1994; Huelsenbeck and Kirkpatrick 1996), but those statistics were not intended to be maximally sensitive to the LBA bias. To further support our claim that bias is responsible for cases where the simple model outperformed the true model, we introduce a quantitative definition of topological bias that can detect LBA, based on unresolved trees (i.e., star phylogenies, including those with different external branch lengths; we also call these star-like trees). When the correct tree is star-like, if a good, unbiased method is forced to choose a best, resolved (i.e. binary branching) topology, then the method will have no alternative but to choose at random

Table 1
Accuracy comparison of true and false models on original and modified trees.

| Tree Name[a] | Tree Specification[b] | Percent correct[c] | |
|---|---|---|---|
| | | True model | False model |
| B | $((.5,.5),.1,(.6,1.4))$ | 72 | 81 |
| B1 | $((.6,.5),.1,(.5,1.4))$ | 77 | 74 |
| C | $((.1,.5),.1,(.2,1.0))$ | 94 | 63 |
| C1 | $((.1,.2),.1,(.5,1.0))$ | 86 | 100 |
| D | $((.05,.05),.05,(.05,.5))$ | 98 | 100 |
| D1 | $((.05,.1),.05,(0,.5))$ | 100 | 90 |
| E | $((.05,.5),.05,(.5,.5))$ | 78 | 80 |
| E1 | $((.05,.55),.05,(.5,.45))$ | 77 | 68 |

[a]Trees B, C, D and E are trees simulated by Yang (1997); B1, C1, D1 and E1 are trees we feel are similar but will be treated differently by bias.
[b]Expected branch lengths, grouped according to topology.
[c]Percent of topology estimates correct using either the true model (JC+G) or the simpler false model (JC), based on 1000 simulations of sequences 2000 bases long, rounded to the nearest percent.

from the possible resolved topologies. There is no reason to choose one topology more often than the others, so in the case of four taxa, each of the three topologies should be chosen in one-third of the simulations. Thus, for a 4-taxon unresolved tree, we may quantitatively define the topological bias of a method in favor of any one topology as the fraction of simulations in which that topology is chosen, minus the expected one-third.

This definition can be extended to resolved trees (such as Yang's) by applying it to an unresolved tree that is as similar as possible to the given tree. As a way of defining the most similar tree, let the "most likely unresolved tree" (MLUT) be the star-like tree that would be found by simulating infinitely long sequences on the original tree and performing ML tree reconstruction on these sequences holding the internal branch fixed at length zero, using the true model for both simulation and reconstruction.

Table 2 shows the biases that result from applying ML using the JC+G and JC models to reconstruct trees from data generated under JC+G on the MLUTs of Yang's trees: tree B0 is the MLUT of tree B, tree C0 the MLUT of tree C and so on. A positive bias means that the original topology was reconstructed more than one-third of the time, indicating a bias in favor of the original Yang tree.

In each case presented in Table 2, the false model is significantly ($p < .01$) biased. The true model is much less biased and its results are consistent with zero bias, except for case C0, where the true model demonstrates a small but significant LBA bias. The bias of the false model favors trees B, D and E (positive bias), for which Yang found the false model to outperform the true model, but disfavors tree C (negative bias), for which Yang found the true model to be better. The amplitudes of the biases are sufficient to be a believable explanation

Table 2
Bias comparison of true and false models on MLUTs of original trees.

| Tree Name | Tree Specification | Percent Bias[a] | |
|---|---|---|---|
| | | True model | False model |
| B0 | $((.512, .512), 0, (.680, 1.460))$ | $-2$ | 14 |
| C0 | $((.503, .118), 0, (.281, 1.052))$ | $-4$ | $-25$ |
| D0 | $((.0503, .0503), 0, (.0998, .5384))$ | 1 | 51 |
| E0 | $((.0604, .4975), 0, (.5347, .5347))$ | 1 | 5 |

[a]Bias defined as percent of topology estimates "correct" (despite zero internal branch length), minus $33^1/_3\%$. Expected standard deviation based on the 1000 trials is 1.5%.

for the differences found in Table 1.

Strikingly, the bias in the simple model gets progressively worse for longer sequences. For example, based on 5000 repeated simulations of tree B0, the false model bias goes from 10% to 19% as the sequence length increases from 2000 to 8000; and it can be shown that in the limit of infinitely long sequences the false model always returns the LBA topology. Thus, ML with an over-simple model can be "positively misleading" in the sense of Felsenstein (1978).

A plausible connection between LBA and the bias in trees D and E may be inferred from the presence of two longest branches in both MLUTs D0 and E0, with the observed bias in the direction of joining these longest branches. The MLUTs seem to have allocated some of the internal branch length of the original tree to the various external branches unequally, so that the short pairwise distances between taxa are nearly the same as in the original tree. This makes sense given the smaller confidence interval on the reconstruction of shorter distances (Bulmer 1991).

To summarize, for each of Yang's trees that was better reconstructed by the false model, we could readily find nearby trees that are better reconstructed by the true model. Furthermore, examination of unresolved trees which are similar to the original trees suggests that use of the false model leads to biased estimates of topology, sometimes biased in the right direction (trees B, D and E), sometimes in the wrong direction (tree C), while the true model gives relatively unbiased estimates.

Yang's results with the trees generated by a Yule process may also be explainable in terms of bias, although we have not investigated this directly. Certainly, on trees obeying a strict molecular clock, long branches must be joined; hence such trees are favored by an LBA bias. Although Yang's trees had fluctuations in their branch lengths that caused deviations from the molecular clock, these trees could still have enough clock-like character to be favored by LBA.

We expect that with real, contemporaneous, biological sequences, the LBA bias may be slightly more likely to favor the correct tree than an incorrect tree due to whatever clock-like tendency evolution may possess. Incorporating a clock-like bias into a model for sequence evolution could improve our ability to reconstruct the correct tree. This suggests the importance of research into models that only mildly violate the molecular clock (Thorne, Kishino

and Painter 1998).

Our results raise concerns about the methodology of testing reconstruction methods, since a biased test-set can cause a biased method to appear to outperform a better, less biased method, as Yang indeed found. One lesson we take away from our results is that a choice of trees that seems intuitively fair may not be. The problem of choosing test trees that can be used to compare biased methods to unbiased ones does not have an obvious, robust solution. Highly symmetric trees, such as Yang's tree A (in which all four external branches are of equal length) would have symmetric MLUTs and should be essentially immune from the LBA bias, but such trees alone would not suffice to expose all the possible deficiencies in a method.

The bias problem also has consequences for studies of molecular evolution in general. Since no model can be assumed to be entirely correct for real sequences, trees reconstructed by any method are likely to be biased in some direction. These biases are systematic and may be positively misleading—they will not disappear but instead grow stronger with longer sequences, and can result in consistently high bootstrap values for some incorrect branches. Therefore, bias may lead to false biological conclusions; overly simple/homogeneous models can cause rapidly evolving taxa to be confidently but incorrectly grouped.

Obviously, one would like to be sure that reconstructed phylogenetic groupings are not caused by bias. One approach toward this end would be to show that a grouping is supported by methods covering a range of biases, where that range is large enough so that it likely includes zero bias. Thus, one could check for consistency between results with models that underestimate long distances (such as a model with no rate heterogeneity) and models that overestimate distances (such as a model with an extremely large amount of rate heterogeneity). The latter models will reverse the bias, resulting in a "long branches repel" effect; as an example, we find that the JC+G model with $\alpha = 0.04$ has a bias of $-10\%$ on tree B0 of Table 2). Unless a branch is supported under both types of models, it should be considered suspect.

**Acknowledgments**

We thank Paul O. Lewis for helping with the sequence simulations, and Jeffrey L. Thorne for useful discussions and insight. A. L. H. is supported in part by NIH grant 5P20–RR11830–02, and the Albuquerque High Performance Computing Center. W. J. B. is supported by DOE contract W-7405-ENG-36.

## Literature Cited


BULMER, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Mol. Biol. Evol.* **8**, 868–883.

FELSENSTEIN, J. 1978. Cases in which parsimony and compatibility will be positively misleading. *Syst. Zool.* **27**, 401–410.

GAUT, B. S., and P. O. LEWIS. 1995. Success of maximum likelihood in the four-taxon case. *Mol. Biol. Evol.* **12**, 152–162.

GOJOBORI, T., K. ISHII, and M. NEI. 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J. Mol. Evol.* **18**, 414–423.

HUELSENBECK, J. P., and D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* **42**, 247–264.

HUELSENBECK, J. P., and M. KIRKPATRICK. 1996. Do phylogenetic methods produce trees with biased shapes? *Evolution* **50**, 1418–1424.

KUHNER, M. K., and J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459–468.

RUSSO, C. A. M., N. TAKEZAKI, and M. NEI. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.* **13**, 525–536.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. In Hillis, D. M., C. Moritz, and B. K. Mable, editors, *Molecular Systematics*, page 442. Sinauer Associates, Sunderland, Mass., 2nd edition.

THORNE, J. L., H. KISHINO, and I. S. PAINTER. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* in press.

YANG, Z. 1997. How often do wrong models produce better phylogenies? *Mol. Biol. Evol.* **14**, 105–108.



Reviewing Editor: STANLEY SAWYER